



Pattern Recognition Letters
journal homepage: www.elsevier.com

UAV First View Landmark Localization with Active Reinforcement Learning

Xinran Wang^a, Chao Li^a, Leijian Yu^b, Lirong Han^a, Xiaogang Deng^a, Erfu Yang^b, Peng Ren^{a,**}

^aCollege of Information and Control Engineering, China University of Petroleum (East China), Qingdao 266580, China.

^bDepartment of Design, Manufacture and Engineering Management, University of Strathclyde, Glasgow G1 1XJ, United Kingdom.

ABSTRACT

We present an active reinforcement learning framework for unmanned aerial vehicle (UAV) first view landmark localization. We formulate the problem of landmark localization as that of a Markov decision process and introduce an active landmark-localization network (ALLNet) to address it. The aim of the ALLNet is to locate a bounding box that surrounds the landmark in a first view image sequence. To this end, it is trained in a reinforcement learning fashion. Specifically, it employs support vector machine (SVM) scores on the bounding box patches as rewards and learns the bounding box transformations as actions. Furthermore, each SVM score indicates whether or not the landmark is detected by the bounding box such that it enables the ALLNet to have the capability of judging whether the landmark leaves or re-enters a first view image. Therefore, the operation of the ALLNet is not only dominated by the reinforcement learning process but also supplemented by an active learning motivated manner. Once the landmark is considered to leave the first view image, the ALLNet stops operating until the SVM detects its re-entry to the view. The active reinforcement learning model enables training a robust ALLNet for landmark localization. The experimental results validate the effectiveness of the proposed model for UAV first view landmark localization.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

An unmanned aerial vehicle (UAV) is an aircraft without a human pilot aboard. It is normally operated by human remote control via radio communication or by autonomous control schemes. The UAVs have many advantages especially when they fly under autonomous control schemes. In this scenario, UAVs operate in terms of automatic navigation and are capable of executing tough tasks that are hard for human beings. One requirement for smoothly conducting UAV automatic navigation is autonomous landing. In particular, one challenging scenario is that a UAV is required to autonomously land on a designated location that is possibly featured with some special marks to aid the landing task.

Most existing UAV autonomous landing techniques rely on certain Global Navigation Satellite System (GNSS) technologies such as the Global Positioning System (GPS). In this scenario, it is potentially an effective manner of exploiting so-

phisticated graph models (Bai et al., 2017, 2018a,b) to optimize the communication network between a UAV and satellites for achieving accurate landing. However, the GNSS-oriented landing techniques have limitations which constrain them from broader applications. First, the positioning accuracy provided by most civil GNSS measurements varies from a few meters to a few hundred meters. It is not easy for the civil GNSS to locate a target within one meter. Therefore, the GNSS can only guide a UAV to approximately land at the designated location but cannot guarantee the accuracy of landing location within meters. Second, there are lots of places (e.g. somewhere inside a building) where the GNSS signals are not available. In this scenario, navigation techniques without using GNSS should be considered for guiding the autonomous landing. To address these limitations of the GNSS-oriented autonomous landing, a vision-based strategy for UAV autonomous landing is investigated in this research. The vision based methods have been broadly explored in the literature of remote sensing (Bai et al., 2014, 2015; Wang et al., 2019), to which the UAV landing scenario has certain similarities. Furthermore, the vision-based methods have the capability of locating a target within meters and landing a UAV in narrowly sheltered space where GNSS signals are pos-

^{**}Corresponding author:
e-mail: pengren@upc.edu.cn (Peng Ren)

sibly blocked. It is assumed that the UAV is equipped with an on-board camera which oversees the land. The camera captures first view videos (i.e. image sequences) of the scenes beneath the flying UAV. It is also assumed that there is a landmark sign on the land designating the landing location. **The research aim is to develop an automatic module that catches the landmark and then steadily localizes it in a UAV first view video, for the purpose of guiding the UAV to autonomously land on the landmark.** The key part in the automatic module is the first view landmark localization algorithm, because the low accuracy of landmark localization such as incorrect detection forms a major reason that leads to UAV's landing failure (Yu et al., 2018). It is clear that the UAV first view landmark localization is an instance of object localization in computer vision and pattern recognition literature, and our work is an inter-disciplinary study between computer vision (Zhang et al., 2013; Bai et al., 2018c) and UAV navigation.

Incorporating computer vision technologies into UAV navigations has already enabled many practical applications such as public safety monitoring, post-disaster rescue, etc (Luo et al., 2018). In our work, we focus on developing robust landmark localization algorithms. To this end, the relevant computer vision algorithms are reviewed in the following paragraphs.

Object localization is a fundamental problem in computer vision. This problem is normally considered as that of detecting a target object with a tight bounding box that covers it (Everingham et al., 2010; Gao et al., 2017; Russakovsky et al., 2015; Vora and Raman, 2018). Recently, the methods of using convolution neural networks (CNNs) (Hong et al., 2015; Nam and Han, 2016; Sermanet et al., 2013; Simonyan and Zisserman, 2014; Wang et al., 2015) have attracted a lot of attention in the literature of object detection. Especially, the R-CNN proposed by Girshick et al. (Girshick et al., 2014; Ren et al., 2017) has been validated to have effective performance for detecting objects in complex backgrounds.

However, it is intractable to straightforwardly apply the CNN-based object detection methods to the landmark localization scenario. The CNN-based methods (Hong et al., 2015; Li et al., 2014; Nam and Han, 2016; Wang et al., 2015) require a large amount of labeled training data for conducting supervised learning. However, such big training datasets in terms of UAV first view video are not always available and this confines the straightforward application of the CNNs to UAV autonomous landing. Additionally, the CNN-based methods do not operate in terms of a sequential mechanism and they are normally used to detect objects in still images rather than videos. Applying the repetitive CNNs to individual images in a video would induce expensive computation on redundant data, which is unacceptable for the UAV on-board computation with limited on-board computational resources.

In contrast to the supervised learning fashion for training CNNs, reinforcement learning methods do not optimize a model subject to a large amount of labeled data but train an agent by the interaction with external environments. This is achieved by employing a sequential trial and error mechanism, and the agent accordingly learns the optimal strategy by maximizing the sum of rewards (Sutton and Barto, 1998). The sequential mechanism of reinforcement learning is suitable for processing UAV first view videos in an efficient manner. In the light of these observations, we propose to address the UAV first view landmark localization problem via reinforcement learning. In our previous study, the action-decision networks (Yun et al., 2017) were exploited and a preliminary reinforcement learning model (i.e. the LLNet) was introduced for UAV first view landmark localization (Wang et al., 2018). We make substantial extensions to the LLNet in our present work. The LLNet does not perform robustly in the situation when the landmark is occluded. If the landmark leaves and re-enters the first view image sequence, the LLNet cannot re-locate the landmark. One reason for this deficiency is that the LLNet uses the interaction of union (IoU) to characterize the rewards in reinforcement learning. The IoU scheme accounts for a reasonable score for landmark localization but lacks the reliable capability of searching a big neighborhood for the landmark. **To address this deficiency, we propose to replace the IoU by support vector machines (SVMs) for characterizing rewards in the reinforcement learning. Specifically, to achieve the goal of accurately locating a bounding box that surrounds the landmark in a first view image sequence, we employ SVM scores on patches surrounded by the bounding box as rewards such that the bounding box transformations are learned as actions. Each SVM score indicates whether the landmark is detected by the bounding box and it can thus judge whether the landmark leaves a first view image. The first view images that lose the landmark are no longer processed by the reinforcement learning actions. On the other hand, the SVM keeps operating on these no-landmark images, searching for the landmark. Once the landmark is found to re-enter the first view image sequence, the reinforcement learning actions are re-started for landmark localization. The reinforcement learning actions operate with the aid of the SVM, which informs the whole model about when the target data (i.e. the images containing a complete landmark) become unavailable. In this scenario, the SVM behaves in a manner similar to the basic idea of active learning. However, unlike most active learning schemes which interactively query information source in the training process, our model performs interactions with data in the execution process. Our proposed strategy is not only dominated by reinforcement learning but also supplemented by an active learning motivated fashion. It is a preliminary attempt of active reinforcement learning. We refer to our new model as**

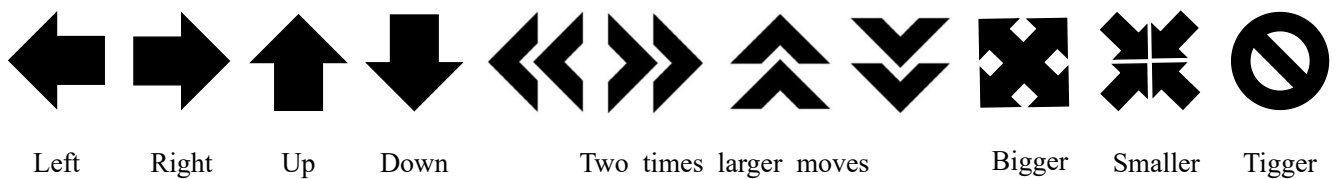


Fig. 1: Action set A.

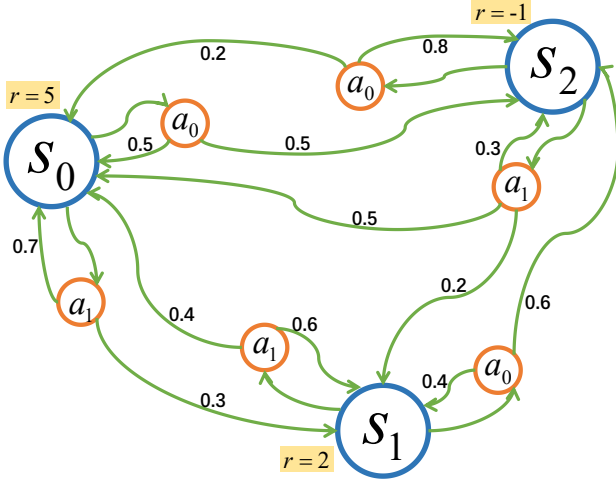


Fig. 2: An example of a Markov decision process with three different states (blue circles), two actions (orange circles), and rewards (yellow shades). The numbers beside the green lines represent the state transition probabilities.

active landmark-localization network (ALLNet). The ALLNet is a robust landmark localization method because it effectively addresses the landmark loss/occlusion situations. Experimental results validate the effectiveness of the proposed method.

The rest of the paper is organized as follows. Section 2 formulates the landmark localization problem as a Markov decision process (MDP). Section 3 presents our model, i.e. the active landmark-localization network (ALLNet), which locates the position of the landmark. Section 4 presents the details about training and executing the ALLNet. Section 5 provides experimental results for comparisons before this research is concluded in Section 6.

2. Landmark Localization as A Markov Decision Process

The landmark localization problem is formulated in terms of a Markov decision process (MDP). An MDP is characterized by five elements, i.e. a finite set S of states, a finite set A of actions, a state transition probability $P_{s,a}(s')$ for each triplet of current state $s \in S$, action $a \in A$ and next state $s' \in S$, a reward functions $r(s)$ for each state $s \in S$, and a discount factor γ . An example of an MDP is shown in Fig. 2, where the subscript for s indicates a time step and the subscript for a indicates an action index. At the time step t , the agent is in the state s_t and the agent chooses the action a_t . At the next time step $t+1$, the agent moves to the next state s_{t+1} subject to the probability $P_{s_t,a_t}(s_{t+1})$. The state transition probability $P_{s,a}(s')$ describes how likely the action a directs the agent from the current state s to the next state s' . At each state, the agent receives a reward $r(s)$. The MDP is a fundamental strategy for describing reinforcement learning algorithms. It provides a formulation for decision making problems. One key goal of a reinforcement learning algorithm in terms of the MDP is to train an agent with the sequential interaction of external environments. The agent is a decision-maker that decides the action to move from the current state of the MDP to the next state. The sequential states of the MDP are to a certain degree under the control of the agent.

In our landmark localization model, we consider a bounding box that possibly surrounds the landmark in a UAV first view image as the agent. The goal of the agent is to locate the

landmark with a bounding box. As shown in Fig. 3, we consider UAV first view images as the environment. The way in which the agent transforms the bounding box for localizing the landmark follows a set of actions. For each image, the agent takes actions until it finally locates the landmark. The agent receives a positive or negative reward at the last state for each image. The value of the reward is determined by a SVM score which judges whether the agent locates the landmark successfully. Specifically, we follow the deep reinforcement learning scheme (Yun et al., 2017) to construct our model. The details of formulating our model in terms of an MDP are presented in the following subsections.

2.1. Transformations of The Landmark Localization Bounding Box as Actions

An action in our model refers to a transformation of the bounding box (i.e. agent). The set of actions A is defined as an eleven dimensional vector. These actions are shown in Fig. 1. Specifically, the actions include four vertical and horizontal actions {left, right, up, down}, their two times larger moves, scale changing actions {bigger, smaller} and the trigger action to stop the locating process. In this way, the bounding box is able to transform in four degrees of freedom.

2.2. Image Patches and History Actions as States

Let (x_t, y_t) denote the center position of the bounding box. Let w_t and l_t denote the width and length of the bounding box, respectively. Let b_t be a 4-dimensional vector and $b_t = [x_t, y_t, w_t, l_t]$. In each frame I_t , the image patch i_t captured by the bounding box is represented as follows:

$$i_t = \lambda(b_t, I_t) \quad (1)$$

where λ is the pre-processing function. In each frame I_t , λ cuts the image patch i_t and resizes it to match the input size of the ALLNet.

Let $h_t \in R^{110}$ denote a binary vector contains the past 10 actions, whose values are set to be zero except the taken action being set to be one. We formulate the state s_t as a tuple (i_t, h_t) .

To enable the state transition, the transformation $f_t(i_t)$ of the image patch i_t is operated as follows:

$$\begin{aligned} x_{t+1} &= x_t + \alpha w_t \\ y_{t+1} &= y_t + \alpha l_t \\ w_{t+1} &= w_t + \alpha w_t \\ l_{t+1} &= l_t + \alpha l_t \end{aligned} \quad (2)$$

where α is empirically set to be 0.03.

Furthermore, the transformation $f_a(h_t)$ of the action history vector h_t is effected by a first-input-first-output (FIFO) scheme and represented as follows:

$$h_{t+1} = f_a(h_t, a_t) \quad (3)$$

The image patch transformation (2) and the history action transformation (3) form the overall transition from the current state to the next state.

We consider the landmark localization process as a deterministic process, in which the next state is determined by the action. Therefore, the MDP for the landmark localization process does not require state transition probabilities.

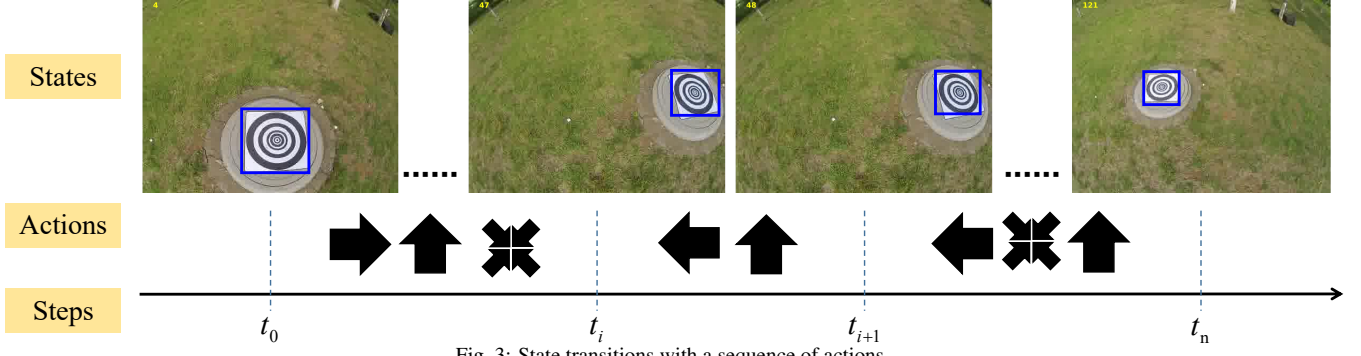


Fig. 3: State transitions with a sequence of actions.

2.3. Support Vector Machine Scores as Rewards

We use support vector machine (SVM) scores as rewards in the MDP model. Here we employ a two-class SVM that classifies image patches into the landmark class labeled as +1 and the background class labeled as -1, separately. We feed the feature $\psi(b_i)$ for the image patch i_t into the SVM and have the soft score c_t^s for classifying i_t as follows:

$$c_t^s = \hat{\omega} \cdot \psi(b_i) \quad (4)$$

Normally, a positive c_t^s means that the image patch is classified into the landmark class. In addition, we define the hard SVM score c_t^h as follows:

$$c_t^h = \text{sign}(c_t^s) \quad (5)$$

We use the hard SVM score c_t^h as the reward $r(s_t)$:

$$r(s_t) = c_t^h \quad (6)$$

The reason that we use the hard SVM scores as the rewards rather than the soft ones is that we will train our ALLNet in a reinforcement learning fashion following the training procedures introduced in (Yun et al., 2017), in which the binary reward scheme renders an effective implementation. On the other hand, the soft SVM scores will be used for judging whether the landmark leaves or re-enters the view.

3. The ALLNet for Performing Landmark Localization

To implement the MDP-oriented landmark localization in terms of a learning strategy, we introduce an active landmark-localization network (ALLNet), which iteratively pursues the position of the landmark. The architecture of the ALLNet is illustrated in Fig. 4.

To solve the MDP problem in our work, small CNN models are more effective compared with deep models (Nam and Han, 2016). Though we define the rewards in terms of the SVM scores, which are different from the rewards used in the action-decision network (Yun et al., 2017), we still follow the action-decision network for constructing our model. First, the action-decision network provides a state of the art reinforcement learning model that is both effective and efficient for practical vision applications. Second, the difference between our ALLNet and the action-decision networks in terms of rewards does not affect the network structure, because rewards normally play a role of the criterion in evaluating the models but do not intrinsically affect the model structures. Specifically, we use the pre-trained VGG-M (Chatfield et al., 2014), a small CNN model, to initialize our ALLNet. As shown in Fig. 4, the ALLNet has three convolutional layers, {conv1, conv2, conv3}. {fc4, fc5} are the next two fully connected layers. {fc4, fc5} both have 512 output units and are combined with ReLU and dropout layers. The output of the fc5 layer is concatenated with the action history vector h_t . The fc6 layer predicts the action probability, which consists of 11 output units. fc7 outputs a confidence which is characterized by the SVM output and measures the probability of the landmark being localized.

4. Training and Execution of the ALLNet

4.1. Training The SVM and ALLNet

The SVM classifies an image patch surrounded by the bounding box into either the landmark class or the background class. To train the two-class SVM, we prepare a number of normalized image patches. Half of them cover the landmark, and the

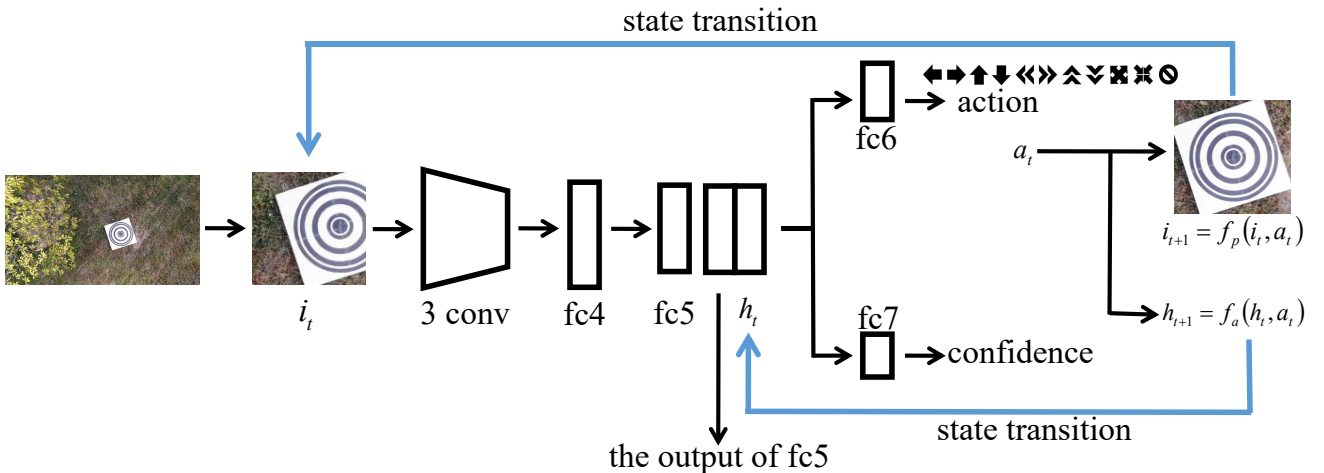


Fig. 4: Architecture of the proposed ALLNet

other half are various background patches with no landmark. We train the SVM as follows:

$$\hat{\omega} = \operatorname{argmax}_{\omega} \frac{1}{2} \omega \cdot \omega + C \sum_{t=1}^T [1 - c_t^h \cdot \omega \cdot \phi_t(b_t), 0] \quad (7)$$

where C is a regularization parameter and T is the total number of patches used in training the SVM. We employ the data augmentation strategy, which is widely used in training remote sensing image classifiers (Yu et al., 2017), to enlarge the training dataset. New training samples are generated in terms of transforming landmark patches with respect to more vast rotations, scales, viewing angles, etc. The use of the augmented training dataset makes the SVM and ALLNet more robust to the landmark variation with respect to rotations, scales, viewing angles, etc.

We follow the training strategy for the action-decision network (Yun et al., 2017) to train the ALLNet. The only difference is that we use the hard SVM scores as rewards rather than the IoU characterized rewards.

4.2. Alternating Executions of The ALLNet and SVM

In an ideal situation where the landmark is fully included in the first view image with no occlusions, we use the ALLNet to locate the landmark. For the same image, the SVM scores the bounding box according to (4). Normally, a positive soft SVM score means that the image patch is classified into the landmark class. More specifically, a large soft SVM score indicates a big confidence of classifying the image patch into the landmark class. In our work, considering the possible partial occlusion of the landmark, we empirically set a warning threshold η which is a positive value smaller than one. Once the soft SVM score is smaller than η , the agent warns that the landmark may be (partially) lost or occluded. In other words, a soft SVM score larger than η reflects that it agrees with the ALLNet landmark localization result. In this scenario, the ALLNet plays a dominant role in landmark localization and the soft SVM score is only computed once for each state. However, once the SVM produces a soft score smaller than η for a UAV first view image, it judges that the landmark is not (fully) captured and the ALLNet (partially) loses the landmark.

There are several reasons for the ALLNet to lose the landmark. The landmark may leave the view and may also be occluded by something else in the view. It may also be due to that the ALLNet does not correctly capture the landmark which still appears in the view. No matter what reason it is for the landmark loss, at the time when the landmark is judged to be lost, the ALLNet stops operating and the SVM takes charge. In this scenario, the SVM searches the whole image for the landmark by placing the bounding box at a position with the largest positive score. In the condition that the largest positive score is smaller than η , the bounding box thus obtained is considered to cover a partially occluded landmark. If there is no positive SVM score over the whole image, this means the landmark is totally lost. The SVM proceeds to process the next image by searching it for landmark and keeps this operation image by image until a patch with a positive soft score larger than η is found in one image. Then the SVM locates the largest positive score patch which is used as the current state to re-start

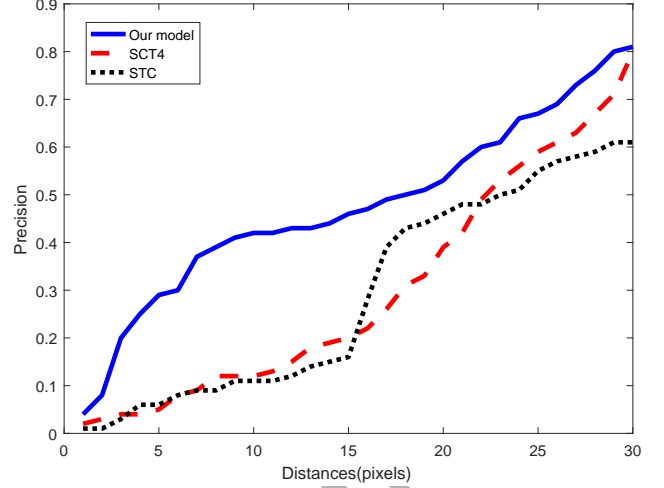


Fig. 5: Percentage of frames with respect to the pixel distance between the center position of the located landmark and that of the ground truth.

the ALLNet. At this time, the SVM stops searching the whole images for the landmark and the ALLNet takes charge for the landmark localization.

The alternating executions of the ALLNet and the SVM renders an efficient and robust landmark localization strategy. The ALLNet is modeled in small size and only processes the local patches surrounded by the bounding box. Therefore, it is computationally efficient. However, it does not account for landmark losses. The operations of using the SVM to search whole images for the landmark are effective to re-locate the landmark. The trade-off is that the SVM operating over whole images takes more computational resources than the ALLNet. The alternating executions of the ALLNet and the SVM balance the model robustness and computational efficiency, and thus enable a practical strategy for the UAV first view landmark localization.

One additional observation is that the ALLNet operates with the aid of SVM such that the SVM score indicates when the output of the ALLNet fails the task and new sample patches are needed. This is very similar to the basic idea of the active learning that is able to interactively query information source to obtain the desired outputs at new data points. In this scenario, our ALLNet can be considered as an active learning motivated model. Therefore, our work is a preliminary attempt of developing an active reinforcement learning strategy for UAV first view landmark localization.

5. Experimental Results

We record a series of UAV first view downward looking videos for evaluating our model. We use VOT2016 videos (Kristan et al., 2015) and a part of our recorded videos to train the proposed model. We use parts of our recorded video that are not the same as those used in training to validate the proposed model.

5.1. Qualitative Evaluations

We validate our method on two different scenarios. The first scenario is that the landmark always appears in the view. The



Fig. 6: UAV landmark localization results from different heights and rotations. Our model locates the landmark successfully in all testing frames.

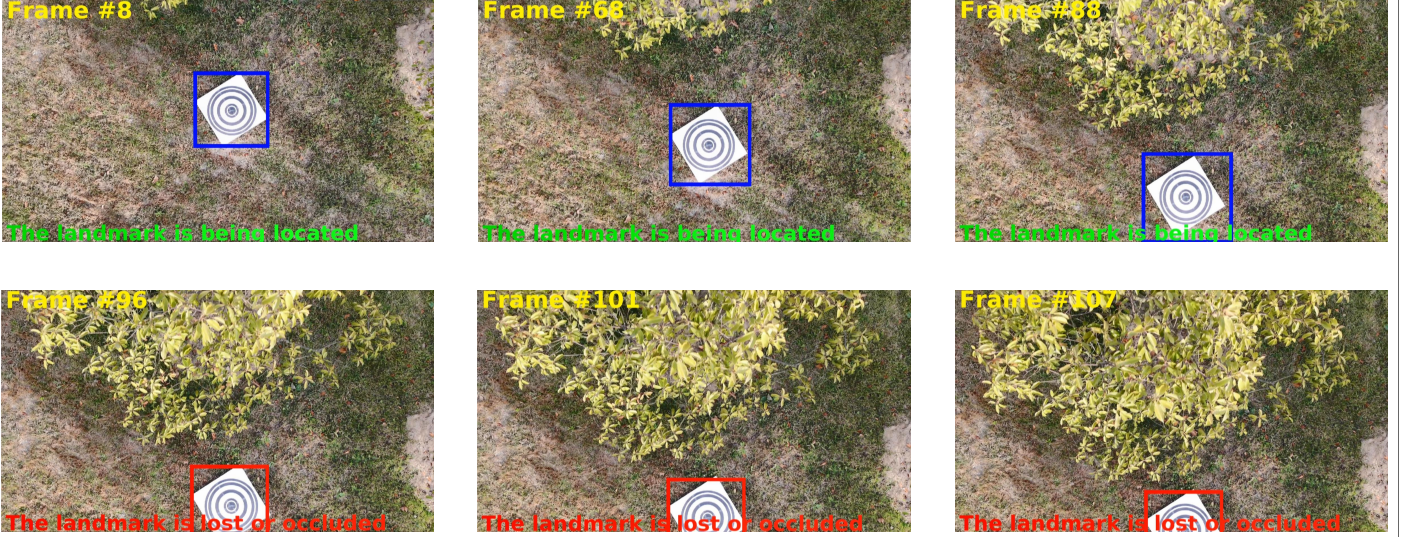


Fig. 7: A scenario that the landmark will be occluded or leave the view. At the beginning, the landmark is completely viewed, and our model is able to locate the landmark precisely. When the landmark starts being occluded and gradually leaving the view, the landmark is precisely predicted, and the agent warns that the landmark is being lost or occluded.

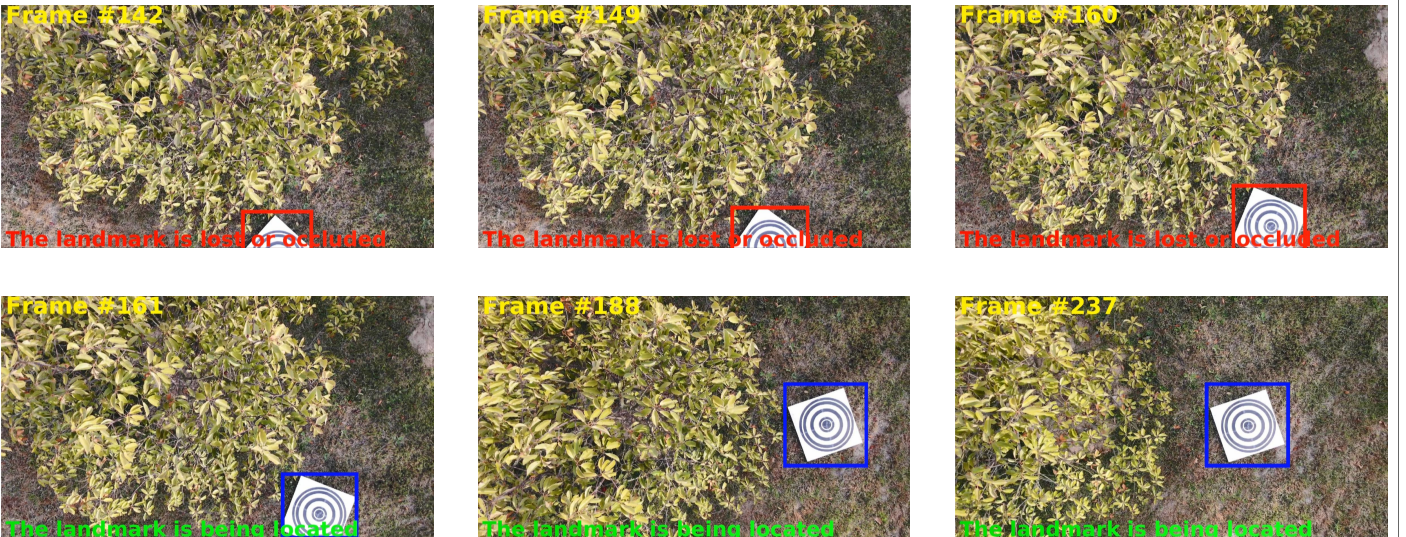


Fig. 8: When the landmark starts re-entering the view, the SVM detects the partially appeared landmark. After the landmark totally appears in the view, the ALLNet locates the landmark precisely.

second scenario is that the landmark leaves the view and later re-enters the view.

The experiment on the first scenario is shown in Fig. 6, in which our ALLNet is able to accurately localize the landmark in all testing frames.

We test our ALLNet on the second scenario. As shown in Fig. 7 and Fig. 8, the ALLNet accurately locates the landmark before it leaves the view and after it re-enters the view. Besides, the ALLNet warns and predicts the landmark precisely both when the landmark is leaving the view and re-entering the view.

5.2. Quantitative Evaluations

We compare our model with two state of the art tracking methods, i.e. STC (Zhang et al., 2014) and SCT4 (Choi et al., 2016), in landmark localization. To make a fair comparison, we use a simplified model of the proposed ALLNet, i.e. the LLNet which is introduced in our previous work (Wang et al., 2018), in this part of comparison experiments. In Fig. 5, we show the percentage of frames with respect to the pixel distance between the center position of the located landmark and that of the ground truth. The results indicate that even the simplified model of our ALLNet is more precise than the comparison methods. At the range from 0 to 30 pixels, we compute the distance between the center of the located landmark and that of the ground truth. Our model has significantly higher precision than the STC and the SCT4 at all time. Specifically, all through our approach, over 80% of the testing frames' errors are within 30 pixels.

To validate the robustness of our ALLNet, we evaluate the precision of the ALLNet for landmark localization before the landmark leaves the view and after the landmark re-enters the view in Table 1. Here the mean center distance error (MCDE) represents the mean pixel distance between the center position of the located landmark and that of the ground truth. The results in Table 1 reflect that the localization errors are controlled within a small range in the situations before the landmark leaving and after landmark re-entry.

Furthermore, we compare the overlap ratio among our model, STC and SCT4. The overlap ratio measures the Intersection-over-Union (IoU) ratio between the located bounding box and the ground truth in each frame. As shown in Fig. 9, the overlap ratio obtained from our model is better than the other two models.

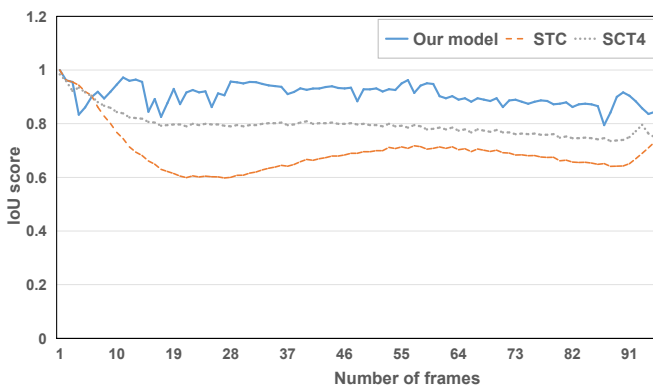


Fig. 9: The overlap ratios of among our model, STC and SCT4 .

Table 1: Localization errors before the landmark leaves the view and after it re-enters the view.

Precision(px)	Before leave	After re-enter
5	13.13%	15.96%
10	55.56%	53.19%
15	87.88%	82.98%
20	93.94%	95.74%
MCDE	11.09	13.55

6. Conclusion

In this paper, we have proposed an effective active reinforcement learning method, i.e. the active landmark-localization network (ALLNet), to address the landmark localization problem. Our method has the ability to locate the landmark even when the landmark leaves and then re-enters the view. We have demonstrated that the reinforcement learning is an efficient scheme to train an agent to learn the landmark localization strategy. This strategy allows the agent to learn from its own history and environments and find the best action policy to locate the landmark precisely. In addition, we have used an active learning motivated strategy to improve our model, making it capable of judging whether the landmark leaves or re-enters the view. This renders a preliminary attempt of combining active learning and reinforcement learning for vision tasks. The experiments have validated the effectiveness of our proposed method.

Acknowledgments

E. Yang is supported in part by the UK Oil and Gas Technology Centre (OGTC) under the LOCUST research project (2019-2021, Grant No.: AI-P-028). L. Yu is funded by the China Scholar Council and the International Fees Only Studentship from the University of Strathclyde (2018-2021). This research is also supported in part under the RSE-NNSFC Joint Project (2017-2019) (Grant No.: 6161101383) with China University of Petroleum (East China).

References

- Bai, L., Cui, L., Bai, X., R. Hancock, E., 2018a. Deep depth-based representations of graphs through deep learning networks. *Neurocomputing*, in press.
- Bai, L., Cui, L., Rossi, L., Xu, L., Bai, X., R. Hancock, E., 2018b. Local-global nested graph kernels using nested complexity traces. *Pattern Recognition Letters*, in press.
- Bai, L., Rossi, L., Cui, L., Zhang, Z., Ren, P., Bai, X., R. Hancock, E., 2017. Quantum kernels for unattributed graphs using discrete-time quantum walks. *Pattern Recognition Letters* 87, 96–103.
- Bai, X., Liu, C., Ren, P., Zhou, J., Zhao, H., Su, Y., 2015. Object classification via feature fusion based marginalized kernels. *IEEE Geoscience and Remote Sensing Letters* 12, 8–12.
- Bai, X., Yan, C., Yang, H., Bai, L., Zhou, J., R. Hancock, E., 2018c. Adaptive hash retrieval with kernel based similarity. *Pattern Recognition* 75, 136–148.
- Bai, X., Zhang, H., Zhou, J., 2014. Vhr object detection based on structural feature extraction and query expansion. *IEEE Transactions on Geoscience and Remote Sensing* 52, 6508–6520.
- Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Return of the devil in the details: Delving deep into convolutional nets, in: *British Machine Vision Conference*.
- Choi, J., Chang, H.J., Jeong, J., Demiris, Y., Jin, Y.C., 2016. Visual tracking using attention-modulated disintegration and integration, in: *IEEE Conference on Computer Vision and Pattern Recognition*.
- Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision* 88, 303–338.

- Gao, F., You, J., Wang, J., Sun, J., Yang, E., Zhou, H., 2017. A novel target detection method for sar images based on shadow proposal and saliency analysis. *Neurocomputing* 267, 220–231.
- Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation, in: *IEEE Conference on Computer Vision and Pattern Recognition*.
- Hong, S., You, T., Kwak, S., Han, B., 2015. Online tracking by learning discriminative saliency map with convolutional neural network, in: *International Conference on Machine Learning*.
- Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Cehovin, L., Fernandez, G., Vojir, T., Hager, G., Nebehay, G., Pflugfelder, R., 2015. The visual object tracking vot2015 challenge results, in: *IEEE International Conference on Computer Vision*.
- Li, H., Li, Y., Porikli, F., 2014. Robust online visual tracking with a single convolutional neural network, in: *Asian Conference on Computer Vision*.
- Luo, C., Yu, L., Ren, P., 2018. A vision-aided approach to perching a bio-inspired unmanned aerial vehicle. *IEEE Transactions on Industrial Electronics* 65, 3976–3984.
- Nam, H., Han, B., 2016. Learning multi-domain convolutional neural networks for visual tracking, in: *IEEE Conference on Computer Vision and Pattern Recognition*.
- Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 1137–1149.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 211–252.
- Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y., 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks, in: *International Conference on Learning Representations*.
- Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition, in: *International Conference on Learning Representations*.
- Sutton, R.S., Barto, A.G., 1998. Introduction to reinforcement learning.
- Vora, A., Raman, S., 2018. Iterative spectral clustering for unsupervised object localization. *Pattern Recognition Letters* 106, 27–32.
- Wang, C., Bai, X., Wang, S., Zhou, J., Ren, P., 2019. Multiscale visual attention networks for object detection in vhr remote sensing images. *IEEE Geoscience and Remote Sensing Letters* 16, 310–314.
- Wang, N., Li, S., Gupta, A., Yeung, D.Y., 2015. Transferring rich feature hierarchies for robust visual tracking. *arXiv preprint arXiv:1501.04587*.
- Wang, X., Ren, P., Yu, L., Han, L., Deng, X., 2018. Uav first view landmark localization via deep reinforcement learning, in: *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition and Structural and Syntactic Pattern Recognition*.
- Yu, L., Luo, C., Yu, X., Jiang, X., Yang, E., Luo, C., Ren, P., 2018. Deep learning for vision-based micro aerial vehicle autonomous landing. *International Journal of Micro Air Vehicles* 10, 171–185.
- Yu, X., Wu, X., Luo, C., Ren, P., 2017. Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework. *GIScience and Remote Sensing* 54, 1–18.
- Yun, S., Choi, J., Yoo, Y., Yun, K., Jin, Y.C., 2017. Action-decision networks for visual tracking with deep reinforcement learning, in: *IEEE Conference on Computer Vision and Pattern Recognition*.
- Zhang, H., Bai, X., Zhou, J., Cheng, J., Zhao, H., 2013. Object detection via structural feature selection and shape model. *IEEE transactions on image processing* 22, 4984–4995.
- Zhang, K., Zhang, L., Liu, Q., Zhang, D., Yang, M.H., 2014. Fast visual tracking via dense spatio-temporal context learning, in: *European Conference on Computer Vision*.